A MULTIMODAL APPROACH TO EXTRACT OPTIMIZED AUDIO FEATURES FOR SPEAKER DETECTION

Patricia Besson, Murat Kunt, Torsten Butz and Jean-Philippe Thiran

Signal Processing Institute (ITS), Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
{patricia.besson,murat.kunt,jp.thiran}@epfl.ch, torsten.butz@a3.epfl.ch
http://itswww.epfl.ch

ABSTRACT

We present a method that exploits the information theoretic framework described in [1] to extract optimal audio features with respect to the video features. A simple measure of mutual information between the resulting audio features and the video ones allows to detect the active speaker among different candidates. The results show that our method is able to exploit the shared speech information contained in audio and video signals to recover their common source.

1. INTRODUCTION

With the increasing capacities of nowadays computers, both auditive and visual modalities of the speech signal can be used to improve speaker detection. Such a detection could lead to great improvements of the user-friendliness of several man-machine interactions. Let us just consider for example a videoconference system. For a proper job, presently one needs an audio engineer and a cameraman so that the speaking person can be emphasized both on audio and video. An intelligent system able to detect the speaker of interest on the basis of sound and image information could focus a moving camera on her/him.

Among the different methods that exploit the information contained in each modality, a few are performing the fusion directly at the feature level. It has been pointed out in [1] and [2] for example, that such a fusion can greatly help the classification task: the richer and the more representative the features, the more efficient the classifier.

Some audio-video feature fusion approaches try to directly evaluate the synchrony of the two signals [3], [4]. As suggested in [4], the synchrony is here the perceptive effect of the causal relationship between the two signals. Other methods map first the features into a subspace where this relationship is enhanced and can therefore be estimated [2], [5], [6]. All the approaches rely on explicit or implicit use of mutual information. An estimation of the features'probability density function (pdf) is therefore required. Normal distributions are often assumed. However, such an *a priori* assumption is not necessarily valid. Fisher in [2], as well as Butz in [1], estimate the probability density functions directly from the available samples during the feature extraction process through Parzen windowing.

Following Butz in [1] and [7], we present here an information theoretic approach to optimize the audio features with respect to video features. The purpose of this method is to detect the current speaker in a video sequence with two or more potential candidates.

The paper is organized as follows: we first present briefly how information theory can be used to extract optimized features in a general multimodal classification problem. We then describe the chosen representation for the video and audio signals. In the third section, the information theoretic optimization approach is applied to obtain audio features optimized for the specific classification task, regardless the classifier. The last part of the paper is dedicated do experiments and discussion about the ability of the method to produce audio features specific to speech leading to speaker detection.

$$\begin{array}{ccccccccc} O & \rightarrow & A & \rightarrow & F_A & \rightarrow & \hat{F}_V & \rightarrow & \hat{V} & \rightarrow & \hat{O} \\ & & & & & \searrow & & & & & \\ O & \rightarrow & V & \rightarrow & F_V & \rightarrow & \hat{F}_A & \rightarrow & \hat{A} & \rightarrow & \hat{O} \end{array}$$

Figure 1: Graphical representation of the coupled Markov chains modelling the multimodal classification process.

2. THEORETICAL FRAMEWORK

The detection of the current speaker in an audio-video sequence can be understood as a classification problem. For a speaker, the audio and video signals originate from the same physical source. Let O be a binary random variable to model the membership to the "speaker" or "non-speaker" class with respect to the observed audio and video signals A and V. Notice that, since no assumption is made on the speaker's position, O can be considered as uniformly distributed. Let F_A and F_V be the features extracted (or mapped) from A and V respectively. They will be viewed as random variables hereafter. Then the estimates $\hat{F}_A$ and $\hat{F}_V$ can be obtained jointly from F_V and F_A by using their joint probability estimation. Then the classification process can be modelled by two first order coupled Markov Chains [1] shown in Fig. (1).

The goal in such a process is obviously to minimize the probability of assigning the measurement to the wrong class. That is, to minimize the classification error probability $P_e = P(\hat{O} \neq O)$ associated to each Markov Chain. Using Fano's inequality and Shannon's entropy, a lower bound on the classification error P_e can be defined for each Markov Chain [1]:

$$P_{e1} \geqslant 1 - \frac{I(F_A, \hat{F}_V) + 1}{\log |\Omega_O|} \quad \text{and} \quad P_{e2} \geqslant 1 - \frac{I(F_V, \hat{F}_A) + 1}{\log |\Omega_O|}, \quad (1)$$

where I is the Shannon's mutual information and $|\Omega_O|$, the cardinality of O which is supposed to remain constant (only two classes in all cases). The best achievable classification error probability P_e is conditioned by the maximization of the mutual information between F_A and $\hat{F}_V$ or F_V and $\hat{F}_A$.

Because of the symmetry property of mutual information, the bounds on the classification error associated to each Markov Chain are equivalent and a joint lower bound can be defined as follow :

$$P_{\{e1,e2\}} \geqslant 1 - \frac{I(F_A,F_V)+1}{\log|\Omega_O|}. \quad (2)$$

Minimizing $P_{\{e1,e2\}}$ comes then to maximize the mutual information between the extracted features F_A and F_V corresponding to each modality.

For maximum mutual information, mapping the features A and V to subspaces F_A and F_V not only reduces the dimensionality of the feature space, but also minimizes the lower bound on the classification error. Moreover, the resulting feature sets can be expected to compactly describe the relationship between the two modalities. The extraction stage, therefore, produces optimized features.

As $H(F_A,F_V) = H(F_A|F_V) + H(F_V|F_A) + I(F_A,F_V)$, it is important to limit possible augmentations of the conditional entropies $H(F_A|F_V)$ and $H(F_V|F_A)$. Indeed, if the entropies increase, they reduce the interfeature dependencies. Dividing Eq. (2) by the joint entropy $H(F_A,F_V)$, a feature efficiency coefficient [1] can be defined as:

$$e(F_A,F_V) = \frac{I(F_A,F_V)}{H(F_A,F_V)} \in [0,1], \quad (3)$$

where again F_A and F_V denote any pair of random variables.

Thus, maximizing $e(F_A,F_V)$ still minimizes the lower bound on the error probability while constraining interfeature independencies.

3. SIGNAL REPRESENTATION

3.1 Video representation

The first processing step consists in choosing a tractable and suitable representation of both signals.

It has been shown in [7] that the audio signal is more related to the pixel intensity changes than to the raw pixel intensities themselves. The video features are thus extracted using the Horn and Schunck's gradient-based algorithm [8] to have a local (pixel-based) representation. The method is implemented in a two frames simple forward difference scheme so that the temporal resolution is large enough to capture complex and quickly varying mouth motions. First a median pre-filtering is used to reduce the noise level. To deal with the curse of dimensionality, only the magnitude of the optical flow and the sign of the vertical component are kept.

Optical flow is computed in each couple of frames over a region of $N \times M$ pixels including the lips and the chin of each speaker. These regions are referred to as mouth regions. Speakers are observed over a sequence of T frames resulting in $T-1$ video feature vectors F_V. The norm of these vectors is normalized to the range [0,1] for the subsequent optimization.

3.2 Audio representation

The audio signal also needs to be represented in a tractable way. This representation should describe salient aspects of the speech signal, preferably similar to those used by the human auditory system, while being robust to variations in speaker or acquisition conditions. Mel-cepstrum analysis is one of the methods that better approaches these requirements and as such, is widely used in speech-processing research [9], [10]. Finally, the speech signal is represented as a set of $T-1$ vectors $\vec{C}$, each containing P mel-cepstrum coefficients $\{C_i(t)\}_{i=1:\ldots:P}$ with $t = 1,\ldots,T-1$ (the first coefficient has been discarded as it pertains to the energy).

4. EXTRACTION OF OPTIMIZED SPEECH AUDIO FEATURES

4.1 Audio feature optimization

In principle, the information theoretic feature extraction discussed in Sec. 2 can now be used for audio and video features F_A and F_V. However, over $T-1$ frames, the dimensionality of the audio features is still too high to be efficiently tractable. It can be reduced in the following way.

For a given set of P weights α_i in a vector $\vec{\alpha}$, an audio random variable $F_A(\vec{\alpha})$ is defined as the linear combination of the mel-cepstrum coefficients:

$$F_A(\vec{\alpha}) = \sum_{i=1}^{P} \alpha_i \cdot C_i, \quad (4)$$

with the weights α_i chosen such that $\sum_{i=1}^{P} \alpha_i = 1$ and $\vec{\alpha} \geqslant 0$). Thus, the set of $P\cdot(T-1)$ parameters is reduced to $T-1$ values $F_A(\vec{\alpha})$. The minimization of the classification error given by Eq. (2) will lead to the optimum vector $\vec{\alpha}$. This optimization requires the availability of the joint probability density as well as of the marginal distributions. These distributions are obviously unknown and will be estimated using Parzen windowing with:

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} h(y-y_i;\sigma), \quad (5)$$

where h is a kernel function whose variance is controlled by the parameter σ, and n is the number of samples available. A Gaussian kernel $G(\mu_A,\mu_V,\sigma_A,\sigma_V)$ is chosen in our case for its widespread validity. The variances σ_A and σ_V are estimated from the data in a robust way, as described in [11]:

$$\tilde{\sigma} = \left(\frac{4}{3n}\right)^{1/5} \cdot \frac{median\,|y_i - \tilde{v}|}{0.6745}, \quad (6)$$

where n denotes the number of samples and $\tilde{v}$ the median of these samples. This implies in our case that the value for σ_V stays fixed for a given set of video features, while σ_A will adapt to the audio features during the optimization process.

4.2 Optimization criteria

As exposed in Sec. 2, minimizing the classification error comes to maximize the efficiency coefficient considering the audio and video features over a mouth region. *Optimization criterion 1* is then defined as:

$$\begin{aligned} \vec{\alpha}_{opt} &= \arg\max_{\vec{\alpha}}\{I(F_V,F_A(\vec{\alpha}))/H(F_A(\vec{\alpha}))\} \\ &= \arg\max_{\vec{\alpha}}\{e(F_V,F_A(\vec{\alpha}))\}. \end{aligned} \quad (7)$$

Notice that in our case the normalization term considers only the audio entropy as the video feature space remains constant.

The results have motivated the definition of a second criterion involving the two mouth regions together. This criterion is referred to as *optimization criterion 2*. It maximizes the squared difference between the efficiency coefficient computed in each mouth regions (referred to by Ω_M and $\Omega_M^{'}$). Thus the optimization is more constrained. Especially, differences between the marginal distributions of the video features in each region are taken into account. If F_V and $F_V^{'}$ denote the random variables associated to regions Ω_M and $\Omega_M^{'}$ respectively, then the optimization problem is:

$$\vec{\alpha}_{opt} = \arg\max_{\vec{\alpha}} \left\{ [e(F_V, F_A(\vec{\alpha})) - e(F_V^{'}, F_A(\vec{\alpha}))]^2 \right\}. \quad (8)$$

4.3 Optimization algorithm

The optimization itself is performed using the unconstrained Powell's direction set method [12] which is deterministic and does not need the analytical form of the objective function.

To reduce the optimization problem as well as to constrain the solution (since $\vec{\alpha}$ is subject to $\sum_{i=1}^{P} \alpha_i = 1$ and $\vec{\alpha} \geqslant 0$), the objective function is re-formulated through trigonometric relations. Namely, instead of directly looking for the set of $\{\alpha_i\}_{i=1,...,P}$ that maximizes the objective function, a set of $\{w_j\}_{j=1,...,\log_2 P}$ weights is defined. Taking advantage of the trigonometric property of Eq. (9), these $\log_2 P$ weights are then combined to define the P coefficients α. If $\log_2 P$ is not an integer, the power of two immediately superior is considered and the weights α are normalized afterwards.

$$\sin^2(w) + \cos^2(w) = \cos^2(\frac{\pi}{2} - w) + \cos^2(w) = 1 \quad (9)$$

$$\alpha_i = \prod_{k_1=0}^{1} \dots \prod_{k_j=0}^{1} \left[\cos^2(k_1 \frac{\pi}{2} - w_1) \dots \cos^2(k_j \frac{\pi}{2} - w_j) \right] \quad \text{with } j = 1, 2, ..., \log_2(P). \quad (10)$$

Thus, the $\vec{\alpha}$ coefficients still constrain the objective function but the number of parameters to optimize is reduced in a logarithmic way.

5. RESULTS

5.1 Experimental protocol

The purpose of the experiments described here was to evaluate the ability of the proposed information theoretic feature extraction method to produce audio features specific to speech signal. The main reason that justifies the use of such a complex method is the need to relate the motion of the mouth to the speech, avoiding non-speech producing mouth motions or speech originated from people not present in the scene.

Tests have been performed by taking two 4s long temporal windows on two gray-scale audio-video sequences where two persons are present, and face the camera. Only one of them is speaking at a time all along the considered temporal window. Notice that persons of two different genders are considered in the second sequence. Since the sequences are sampled at 25 frames/s, each 4s temporal window contains 100 frames. The two persons present will be called "speakers", as they can both potentially be speaking from the detection algorithm point of view. Also, the four extracted temporal windows will be referred to as sequences 1, 2, 3 and 4.

	Optimization Criterion 1	
Sequence.	F_{As}^{opt}	F_{Ans}^{opt}
1	86.78 %	46.41 %
2	63.35 %	-18.47 %
3	88.46 %	86.83 %
4	67.08 %	-28.81 %

Table 1: Normalized difference between the mutual information computed in the speaker and the non-speaker mouth regions for each of the four test sequences. The columns indicate which audio features were used.

First, each mouth region is manually extracted from each of the 100 frames, resulting in two regions of $N \times M$ pixels, where N and M vary between 22 and 33 pixels, depending on speakers' characteristics and acquisition conditions. Thus the video feature set is composed by the $N \times M \times 99$ values of the optical flow norm at each pixel location over the considered time.

From the audio signal sampled at 44100Hz, a 12 coefficients mel-cepstrum is computed using 23.22ms Hamming windows [9], [10].

Considering each mouth region and its associated video features, the mel-cepstrum coefficients are projected on a new subspace as defined in Sec. 4. In a first time, only the *optimization criterion 1* defined in Sec. 4.2 is applied to analyze the ability of the method to extract specific audio features. The discussion of the results leads to the definition of the more efficient *criterion 2* given by Eq. (8).

5.2 Optimization results

As a result of the optimization, two sets of weights are obtained (one for each mouth region). They give the optimal linear combination of mel-cepstrum coefficients with respect to the optimization criterion. Let us denote them $\vec{\alpha}_s^{opt}$ and $\vec{\alpha}_{ns}^{opt}$, where the indices *s* and *ns* indicate whether these weights result from the optimization performed on the speaking mouth region or on the non-speaking one respectively. Two corresponding audio feature sets derive from these weight sets: F_{As}^{opt} and F_{Ans}^{opt}.

Two pairs of mutual information values can be evaluated between these audio features and the video features in each mouth region. If F_{Vs} denote the video features of the speaking mouth region and F_{Vns} those of the non-speaking one, the two pairs of mutual information are given by:

$$\{I(F_{Vs}, F_{As}^{opt}), I(F_{Vns}, F_{As}^{opt})\}, \quad (11)$$

$$\{I(F_{Vs}, F_{Ans}^{opt}), I(F_{Vns}, F_{Ans}^{opt})\}. \quad (12)$$

What we expect is that the largest value corresponds to the mutual information between the video features of the speaking mouth region and the audio features. Notice that the active speaker can not be detected by simply looking at the objective function's final values in each mouth region. Indeed, it has been observed that the optimization could converge towards a larger maximum on the non-speaking mouth.

The results obtained for the four sequences are summarized in Table (1). Values indicate the normalized difference between the mutual information computed in the speaker and the non-speaker mouth regions. It appears that the results obtained with mutual information scheme of Eq. (11) (column

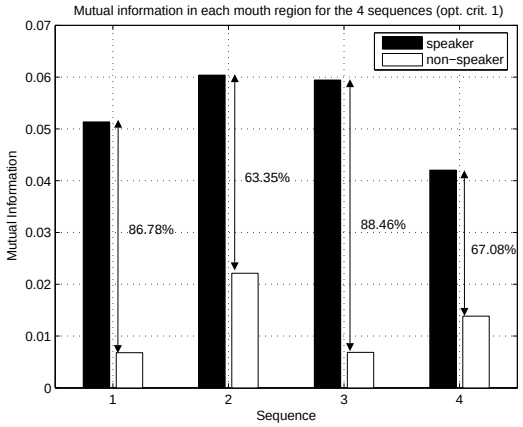

Figure 2: Mutual information obtained in each mouth region for the four sequences when applying scheme of Eq. (11) and using the *optimization criterion 1*.

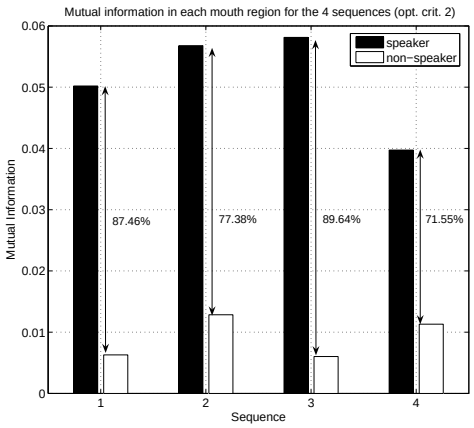

Figure 3: Mutual information obtained in each mouth region for the four sequences using *optimization criterion 2*.

1) are those expected: the right speaker is always pointed out by a larger mutual information value. A graphical representation of these results is shown in Fig. (2). The mutual information scheme of Eq. (12) (column 2) does not indicate the active speaker so clearly, or even fails sometimes (sequences 2 and 4). This is not surprising at all, since the audio features used in that case have been obtained on the non-speaking mouth region. Therefore, even if the optimization algorithm manages to maximize a mutual information based criterion, the output can not (and is not expected to) reflect an underlying relationship between audio and video.

These results show that the method produce audio features with specific information when there exist a relation with video features. This discussion motivated the definition of the second optimization criterion given by Eq. (8) which directly considers the two mouth regions. Thus, only one weight set $\{\alpha_i\}_{i=1,\ldots,P}$ is obtained and resulting optimal audio features can be directly used to measure the mutual information in each mouth region. The results are presented in Tab. (3). The active speaker is correctly detected in each case but this detection is performed in a simpler one-step measure.

6. CONCLUSIONS AND FUTURE WORKS

We have presented a method that exploits the common content of speech audio and video signals to detect the active speaker among different candidates. This method uses the information theoretic framework exposed in [1] to derive optimal audio features with respect to the video ones. No assumption is made about the distributions of the features, rather they are estimated from the samples. The results show that the method is able to extract audio features that are specifically related to the speaker video features. Using only these extracted features, the algorithm performs detection of the current speaker.

Acknowledgements

The authors would like to thank Dr. V. Popovici and K. O'Connor for their contributions and fruitful discussions, as well as the Swiss National Found which supports this work.

REFERENCES

[1] T. Butz and J. P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Processing*, vol. 85, pp. 875–902, 2005.

[2] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transaction on multimedia*, pp. 406–413, 2004.

[3] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *CIVR*, 2003, pp. 488–499.

[4] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *NIPS*, vol. 12, 2000.

[5] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronisation of video facial images and audio tracks," in *NIPS*, vol. 13, 2001.

[6] P. Smaragdis and M. Casey, "Audio/visual independent components," in *ICA*, Nara, Japan, April 2003.

[7] T. Butz and J. P. Thiran, "Feature space mutual information in speech-video sequences," in *ICME*, vol. II, Lausanne, Switzerland, 2002.

[8] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, pp. 185–203, 1981.

[9] B. Gold and N. Morgan, *Speech and audio signal processing*. John Wiley & sons, Inc, 2000.

[10] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proceedings of the IEEE*, vol. 81, no. 9, Sept. 1993.

[11] A. W. Bowman and A. Azzalini, *Applied smoothing techniques for data analysis*. Oxford science publications, 1997.

[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.